DOCUMENT RESUME

ED 053 197 TM 000 728

AUTHOR Ball, Samuel; Bogatz, Gerry Ann
TITLE Summative Research of Sesame Street: Implications for the Study of Preschool-Aged Children.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 71
NOTE 27p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Behavioral Objectives, Child Development, *Cognitive Processes, Data Collection, *Disadvantaged Youth, Early Childhood Education, Educational Television, Interaction, Measurement, Post Testing, *Preschool Children, Pretesting, Research Methodology, *Summative Evaluation, *Symbolic Learning, Testing
IDENTIFIERS *Sesame Street

ABSTRACT

Sixty-six goals, mostly involving symbolic representation and cognitive processes, were evaluated to indicate those aspects of the summative research on Sesame Street that were bases for generalizing about research into the development of preschoolers. Unintended and intended outcomes were investigated as well as the program's effectiveness for various groups of preschoolers. (MS)

ED053197

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

SUMMATIVE RESEARCH OF SESAME STREET: IMPLICATIONS FOR THE STUDY OF PRESCHOOL-AGED CHILDREN

TM 000 728

by

Samuel Ball

and

Gerry Ann Bogatz

Educational Testing Service
Princeton, New Jersey

When a neologism crops up each year to replace the neologism of the year before, one suspects that there are some negative attitudes attached to the referent these new words are supposed to denote. In the distant Australian past, and before coming to Up Over, one remembers the word "foreigner" being replaced by "alien" being replaced by "displaced person" being replaced (through the sure touch of a genius) by the term "new Australian." It is easy to dislike foreigners and aliens. Even displaced persons, while coming in for some pity, can readily be denigrated. But what Australian could hate a new Australian? Well, it took a bit longer but it happened. Neologisms are often euphemisms. And thus today's euphemism can be tomorrow's profanity.

Of course, one does not have to go so far afield for examples of this. When talking to leaders in poverty areas, we were often dismayed to find ourselves talking about the underprivileged, inner-city, core-city, ghetto, dispossessed, culturally-deprived, urban child. And the leaders look at us and say, "I think I know who you're referring to. You mean the so-called underprivileged, inner-city . . . ." And then we get down to business.

So before we get down to business here, let me take up from the title the curious term "summative research." It has gone through some changes lately such as "educational auditing" or "accountability assessment"--two of the more horrible examples

from recent educational jargon. Summative research is itself a neologism or euphemism for "evaluation." It carries, perhaps, more precision than "evaluation" but that precision is in the ear of the sayer anyway. So, why all these new terms for "evaluation"?

Evaluation, especially educational evaluation, has become a shady word especially in the halls of pure science.

> Clearly, evaluative research is an activity surrounded by serious obstacles. Satisfied with informal and impressionistic approaches to evaluation, policy makers are often reluctant to make the investment needed to obtain verifiable data on the effects of their programs. Evaluative researchers are typically confronted with problems of measurement and design which greatly restrict their ability to reach unambiguous conclusions. Abrasive relations with practitioners and clients can add to the evaluator's difficulties in obtaining information. Evaluative research is often addressed to a distressingly narrow range of issues; results are not as fully or widely disclosed as they might be; highly pertinent findings are often ignored by policy makers. It is little wonder that many social scientists regard evaluative research as a dubious enterprise. (Caro, 1971)

In short, too often evaluation is thought of in connection with poor research design, selective perception, imprecision, and oversimplified presentations of hopelessly complex topics. But this does not have to be. I use the term summative research for a special reason,but otherwise I would gladly talk of our evaluation of Sesame Street. I would argue that, properly carried through, educational evaluation can be one of the most fertile sources of data for child psychologists.

If I might interpolate some personal special pleading, I wish that more of the leaders in child psychology would take a more active interest in educational evaluation, not just to improve the science and art of evaluation, but to enable their own research to become, perhaps, less precious.

When Children's Television Workshop began to develop Sesame Street, it felt a need for two research groups. An in-house formative research group whose research and evaluation of segments as they were taped would provide immediate help for the show's producers. And then there was in independent, summative research, out-of-house (sometimes charmingly shortened to out-house) group. Our role was to test the finished product, the first year of Sesame Street, and evaluate it (Ball and Bogatz, 1970). What I hope to do here is to describe what we did and, as I present the description, suggest where implications might be found, at least for those interested in the study of preschool-aged children.

The Goals of Sesame Street

In the summer of 1968, after a series of five meetings, each lasting three days, the goals for the first year of Sesame Street were established. The meetings themselves were innovative, bringing together television writers and producers, educational researchers, Head Start teachers and supervisors, writers and publishers of children's books, librarians, Madison Avenue

advertising executives, movie moguls, psychiatrists, and child psychologists, including some from places as far afield, for example, as Minnesota.

Of major concern to us at ETS was that 66 goals were settled on, mostly couched in behavioral terms. The goals came in four sections:

I. Symbolic Representation (letters, numbers, geometric forms)

II. Cognitive Processes (perceptual discrimination, relational concepts, classification, ordering)

III. The Physical Environment

IV. The Social Environment

Some of the 66 goals were classified as "primary instructional goals" and were the subject of concentrated production efforts. Almost all of these goals received concentrated attention in the evaluation. They were mainly in the cognitive areas involving symbolic representation and cognitive processes.

Research Strategy

Two major principles guided us in the evaluation. First we felt it important to look for unintended as well as intended outcomes. That is, the goals of the show were important, and we certainly hoped to assess the effects of viewing the show in relation to those goals. But we felt that was not enough. The medical model of evaluation reminds us that concentrating on achieving intended outcomes and ignoring side effects can

lead to some horribly wrong overall evaluations--for example, as in the original testing of thalidomide. (Scriven, 1967)

A second major principle we considered was that interactions may tell us more in an evaluation than main effects. That is, in a worthwhile evaluation we must discover not only if the educational intervention, in general, works (an important question, of course). For the long run we should also try to discover which children it works best for, which children it does not seem to work for, and the conditions under which it operates most efficiently. Too often evaluations have concluded that a new program is of little consequence, when in fact it is a boon to some children, a ruin to other children, but when averaged over all children, there seems little difference from the old program.

The application of these two principles in the summative research for Sesame Street caused us to assess at pretest and posttest times not only progress along some 36 primary goals of the show but also transfer effects, home background variables, parental attitudes, and socioeconomic status factors. We decided to sample children from middle class suburbia, lower class northern and western urban ghettos, lower class sections of a southern town, rural children, Spanish-speaking children, children at home and children in Head Start and nursery schools, boys and girls, black children and white children, and 3-, 4-, and 5-year-old children. Initially we tested over 1,300 children. Then we observed many of them viewing the show, made a content analysis of the show itself, administered a questionnaire to teachers whose classes viewed the show, and assessed the amount

of viewing for all the subjects in the study using four different assessment techniques. When evaluating a program in which side-effects and interactions are considered important, the study has to be wide-ranging, the sampling extensive, and the statistics multivariate. (Freeman, 1963)

If this kind of research strategy is reasonable for educational evaluations, it also bears scrutiny for studies of child development generally. A univariate approach to child psychology is reminiscent of the poem of the six blind men examining the elephant by touch and deciding what the elephant must be like on the basis of this partial examination (John Godfrey Saxe). One blind man, for example, felt the tail and said the elephant was like a rope, and another felt the legs and said it was like a tree trunk. This is not to say that small, highly focussed, status or one-occasion studies are not worthwhile but merely to say that the results of such studies must be replicated and, if necessary, reassessed in larger, more comprehensive, longitudinal studies.

Field Research

Sesame Street was primarily intended for preschool-aged, disadvantaged children at home who were without benefit of Head Start or similar educational experience. Therefore, a major thrust in our sampling was to study children who were in this category. Working in ghetto communities is an increasingly difficult problem for researchers. In general, the more militant

the community, the more it looks askance at the clipboard-wielding researcher who comes from outside, studies the community's children and then disappears without any discernible increases in benefits to the children. An evaluator who brings with him a product that might be beneficial to the children is in a potentially more advantageous position than the increasingly distrusted basic researcher, but it is a position that has to be further developed. (Walsh, 1969)

A crucial factor in getting our evaluation work accepted in the days before Sesame Street was known was not our verbal protestations that it was the other fellow who was exploiting their children, nor was it our plea that we wanted to evaluate the show and not evaluate their children. What was crucial was our willingness to appoint local community members as coordinators, testers, and observers. While the income was probably a factor, an important principle seemed to be that the work was being, in a sense, controlled from within. If they were conducting the work, there was less chance that they were being, in some way, hoodwinked. We earnestly recommend that indigenous personnel be employed in developmental studies of low-income children or special groups of any age.

Certain advantages accrued from using indigenous people as our field staff. It meant that many doors were opened to us (both literally and metaphorically) that would not otherwise have been opened. In our house-by-house listing of 3- to 5-year-olds, we had very few refusals (less than 5 percent). Second, at the very least, no harm was done to the validity of our

testing when our testers had accents and dressed and behaved in culturally familiar ways to the preschool-aged subjects.

The third advantage was one we had not counted on initially. Since the testers were not sophisticated in test theory and not advanced educationally themselves, attempts at dishonesty were easily caught when the data arrived at our office.

Problems of honesty in data collection are not new. The literature on the topic seems to indicate that with middle class, well-educated testers and interviewers it is difficult and costly to solve the problem. We had a number of devices built into our test battery to enable quality control to be exercised. We did have to discard the data from four testers and thereby we lost about 130 subjects from our initial 1,300. But the fact that our local coordinators were also indigenous to the community meant that unpleasant supervisory roles could be played without too much fuss and community reaction. Of course, knowing that people with low educational levels would be administering the instruments presented problems when it came to constructing the measuring instruments, but even here it was a blessing disguised as a problem. It meant we had to take a new and patently clear approach to test development, and it is good that we did.

Measurement

The measurement of preschool cognitive knowledge, skills, and processes is usually an esoteric business. Most of us recall taking practicum courses in how to administer tests to young children. Of course, when assessing young children, individual not group tests are appropriate but this has led to a most

unfortunate tradition. The tests, rightly so, are individually administered; however most seem to be laboratory-type instruments involving the tester in subjective judgments and the child in situations which may be rich in clinical insights for the tester but are complicated in terms of generalized meaning.

For the clear assessment of variables, most preschool measures are bad. For example, we might take a toy 'plane and give it to little Johnny and say, "Make the 'plane fly over my arm." And Johnny plays with the 'plane and maybe he makes it fly over my arm but did he mean it to do so, does he understand the term "over", or is he merely a very active child? Or perhaps we are interested in field independence, so we present a hidden triangles test. The child is given a stimulus triangle and asked if he can find one just like that in the picture. We ask him to trace the triangle in the picture with his finger. He cannot? Does this mean he really didn't see it there, does it mean he has poor psychomotor coordination, or does it mean he doesn't want to play games with the tester?

In small studies conducted by experienced testers in laboratory-type situations, these deficiencies of subjectivity and confounded assessment may not be overwhelmingly negative. They reduce reliability and they may affect validity, but they at least provide the tester with some knowledge of individuals and the discipline with a mystique.

When it comes to larger studies in field situations, the problems are magnified. In one longitudinal study involving multiple assessments of poverty children and indigenous community members as testers, the training sessions sometimes took seven or eight weeks. While this speaks highly for the patience and probity of the principal investigator, it says something too about the nature of the measures being used.

In our Sesame Street evaluation we noted that most of the goals to be assessed were behaviorally defined and referred to the cognitive domain. Most of the children we were assessing would be tested in their own homes or in free corners of corridors near classrooms. Kits of toys and complicated procedures were out of the question. Further, we were not primarily interested in obtaining clinical insights into the behavior of individual subjects but rather we were interested in obtaining reliable and valid data on specified groups of children. Finally, as mentioned earlier, most of our testers were relatively uneducated and the nature of our task precluded lengthy training procedures.

With these considerations in mind, Gerry Bogatz, who was in charge of measurement, set to work. She found that with a more than 200 item battery involving two hours of testing (over three or four sessions), our assessment could be accomplished using four basic item types. This simply meant that both the child and the tester could concentrate on the content of the test itself.

All that was required was a set of stimulus pictures. The overt response required of the subject usually was pointing. The child was not required to verbalize unless verbalization was the goal being assessed. The child was not required to interpret the drawings on the stimulus page; he was told what the drawing was supposed to be depicting.

The areas we assessed in the first year included:

Body Parts	Pointing
	Naming
	Functions of
Letters	Recognizing
	Naming
	Matching
	Initial Sounds
	Reading Words
Geometric Forms	Recognizing
	Naming
Numbers	Recognizing
	Naming
	Numerosity
	Counting
	Addition and Subtraction

Matching

Relational Terms — Amount

Size

Position

Sorting

Classification — by Size

Form

Number

Function

Incongruities (Puzzles Test)

Embedded Figures (Hidden Triangles)

Sequencing (Which Comes First)

The median reliability (Cronbach alpha) of the subtest totals at pretest time was .77 (the total score reliability being .96); the median reliability at posttest was .82 (the total score reliability being .98). The tests then were reliable, and they were either clearly keyed to the goals (as in naming of lower and upper case letters), or else measured possible important transfer of learning effects (for example, reading words). It took about two days to train mothers with low educational attainments to administer them. As it turned out the scores were sensitive to the experimental input, and this was a rare, if not unique, event in the educational evaluation of 3- through 5- year old children. In general, tests of 3- through 5-year olds, at least in the cognitive areas, do not have to be complex to administer nor difficult to interpret, though they do need individual administration.

For research and evaluation involving older children and adults the test arsenal contains instruments developed for clinical and laboratory assessment and for more massive evaluational studies. With preschool-aged children the professional arsenal is sadly lacking in instrumentation developed for large-scale field research. Yet much experimental research in child psychology could be made at least more generalizable (to what extent can you generalize from research on 25 children who were spawned by university faculty?)if sturdier, more easily administered, reliable instrumentation were available. Cross-fertilization is needed.

Flexibility and Evaluation

This heading may simply be an instance in which a virtue (flexibility) is made of a necessity. In evaluating Sesame Street, there were a number of instances where, if the original plans had been followed, the study would have been a disaster. It is times like that when those involved in field evaluations look longingly at their more pure colleagues conducting smaller, better controlled types of research on, say, the reaction times of middle class three-year-olds to requests couched in the active and passive voices.

One of the worst problems we had resulted from the unexpected popularity of Sesame Street. We had purposely gone to sites which had VHF rather than UHF ETV stations because we had been worried that too few of our sample would view the show. We had encouraged

some of the children to view and had not encouraged the rest. For this latter, non-encouraged group we had used a pretext for our testing. Some of our advisers had argued eloquently that we ought to pay our encouraged group to view or we would be in the unenviable position of having worked with over a thousand subjects of whom but a handful could be classified as experimental. Fortunately that idea was resisted.

What happened was that the show generated such popularity that only one in eight of our sample failed to view it at all. Self selection swept aside our carefully contrived allocation of subjects to encouraged and non-encouraged conditions. With the inestimable wisdom of hindsight, in the second year's evaluation, we have gone to areas where cable is needed to obtain the show and we have allocated cable to some homes and not to others. This seems to be working well. However, it did nothing to alleviate the problem that arose in our first year study. We found it convenient to say it was now a study of the effects of amount of viewing rather than of viewing versus no viewing, but that was no solution to the basic problem.

It became apparent that the self-selection factor in viewing meant that, using our amount of viewing index, the children who viewed the show most had the highest attainments at pretest. Thus, although the more the children viewed, the more they gained, it was

not clear at first glance to determine whether this was a function of greater viewing or of pre-existing steeper growth rates. We used covariance techniques and found that even with pretest scores, SES, and Peabody scores covaried, viewing was still a significant effect. However, covariance is a controversial, if not erroneous technique in these circumstances. Fortunately there was a better procedure.

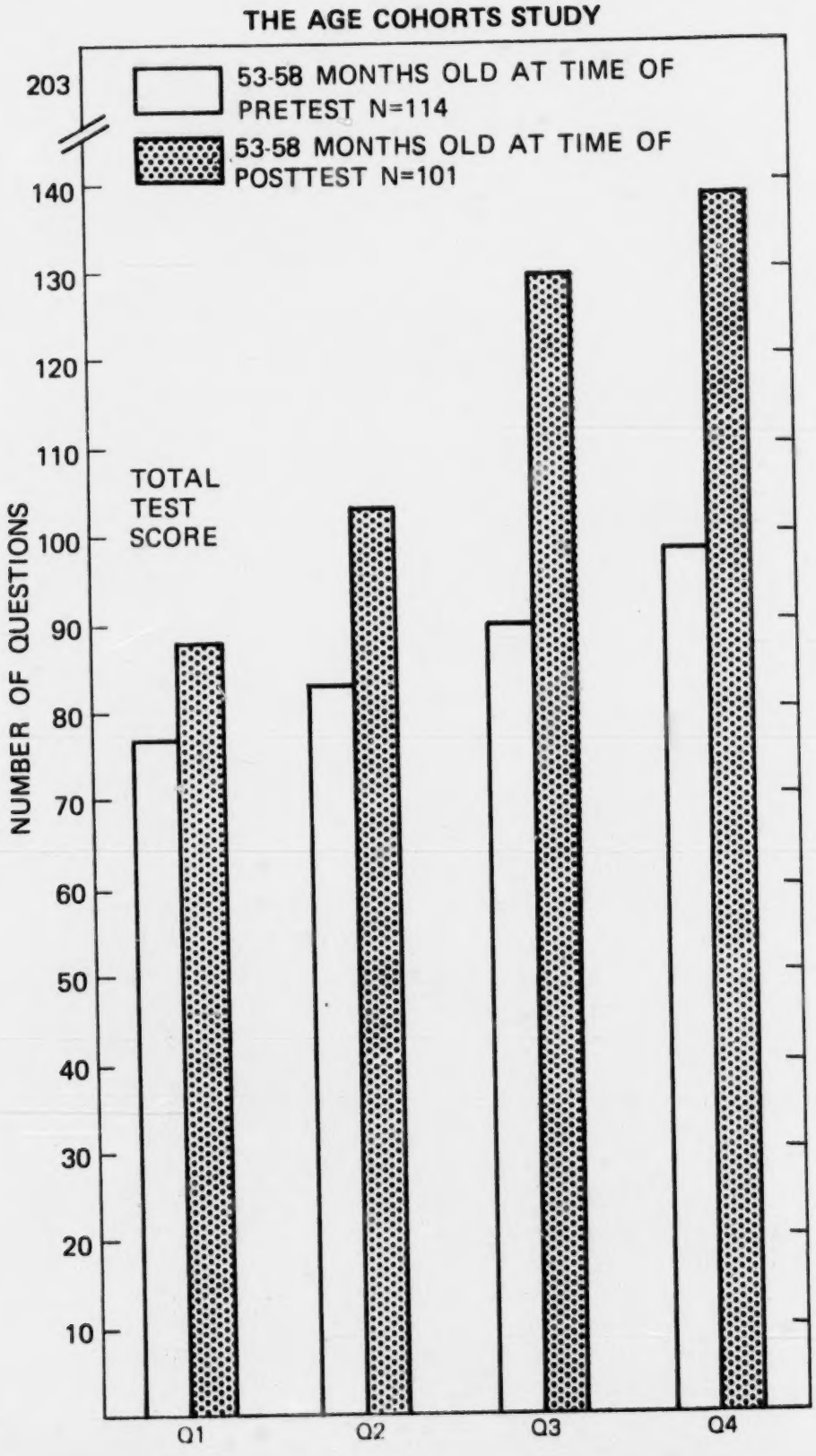
By using pretest scores as a sort of norm group we were able to unconfound the confusion. That is, we took two matched groups of children. Group 1 was 53 to 58 months of age at the time of pretesting; Group 2 was 53 to 58 months of age at the time of post-testing. In addition to being of the same chronological age at the point of comparison, they were of comparable mental age and they lived in the same communities. There were, in short, no observable differences between the two groups in important matters of previous attainments, IQ, and home background. There were more than 100 disadvantaged children who were not attending school in each group. (See Figure 1)

Insert Figure 1 about here

The pretest scores of Group 1 (before the children could have watched Sesame Street) were compared with the posttest scores of Group 2 after the Group 2 children had watched the program. The frequent viewers in Group 2 scored about 40 points higher on the 203 common items than the comparable children in Group 1 who had never watched the show. Equally significant is the fact that the

FIGURE 1

THE AGE COHORTS STUDY



As viewing becomes heavier, (from Q1 to Q4) the achievement differential between "Sesame Street" viewers (shaded) and non-viewing controls (unshaded) shows an increasing advantage in favor of the experimental viewers.

infrequent viewers in Group 2 differed by only about 12 points from comparable children in Group 1 who had not viewed Sesame Street at all. In short, holding maturational effects, IQ, previous attainments and home background constant, the frequent viewers made relatively large and important gains.

Such acts of juggling are essential to action research. They are also useful when we enter new areas of research. Child psychology, at least in some of its areas of research, is not well-advanced. When we are still groping to describe the phenomena of our study, then the flexible, watch-the-data-and-react-to-it approach rather than the more precise, hyphothesis-testing style seems to be the most appropriate one to use. (Rust, 1971) There is a kind of traditional hierarchy in research methodology that seems to put this (flexible or sloppy depending on your viewpoint) approach low in respectability. Our point is simply that if it is appropriate to the situation, it is the best approach to use.

Some Unexpected Results

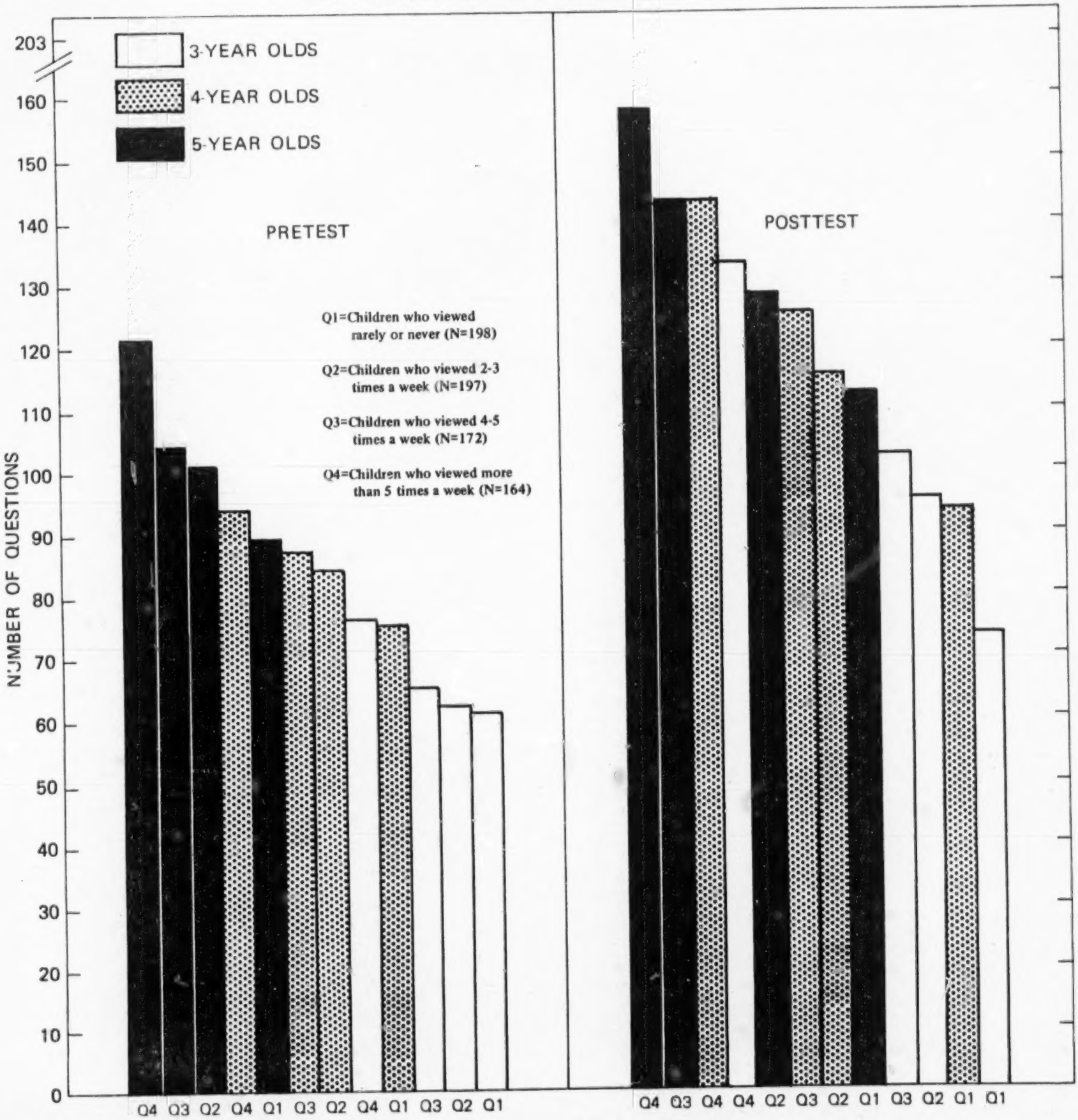
There were three sets of results that were unexpected, at least by us. The first is among the three age groups who watched the show: 3-, 4-, and 5-year-old disadvantaged children. (The word, "disadvantaged" can be, and is, defined in several different ways. We worked in poverty areas and found that economic status, amount of education and attitudes were related factors.)

Insert Figure 2 about here

Figure 2 presents the data on the pretests and posttests given

FIGURE 2

PRETEST AND POSTTEST SCORES OF 3, 4
AND 5-YEAR OLD DISADVANTAGED CHILDREN



Note how at pretest the groups of children are sorted almost entirely by age while at posttest the important factor is whether or not they watched Sesame Street. Note, for example, the high viewing three-year-olds (Q. 4) at pretest as relative performance ranks now in the bottom third , but at post test, this same group is in the top third group, from the eighth rank to the fourth rank.

to all the disadvantaged children. The children were divided into quarters based on the amount they had viewed Sesame Street. Children in Quarter 1 (Q1) viewed never or once a week, Q2 viewed 2-3 times a week, Q3 viewed 4-5 times a week, and Q4 viewed more than 5 times a week. Before Sesame Street went on the air, older children almost invariably performed higher on the test than younger children. After Sesame Street, however, three-year-olds who watched most (Q4) scored higher at posttest than three of the four-year-old groups and two of the five-year-old groups, although these three-year-olds had a pretest score lower than all five-year-olds and all but one of the four-year-old groups.

In other words, the placement of the children along the scale measuring the goals of Sesame Street was very dependent on age at pretest while at posttest it was much more related to amount of viewing. These data also suggest that three- and four-year-olds are capable of learning many of the skills traditionally reserved for the five-year-old in school. And the data also support the general result of the evaluation, namely, that children who watched the most learned the most.

The second unexpected set of results concerns the middle-class four-year-old children in the study and the four-year-old disadvantaged children. Recent history of research has warned that such comparisons are often unwise to make, primarily because so many things differentiate the two groups that a comparison is

likely to be an invidious one, unfairly discriminating against the disadvantaged group. However, in this instance, the comparison is one that should be welcomed.

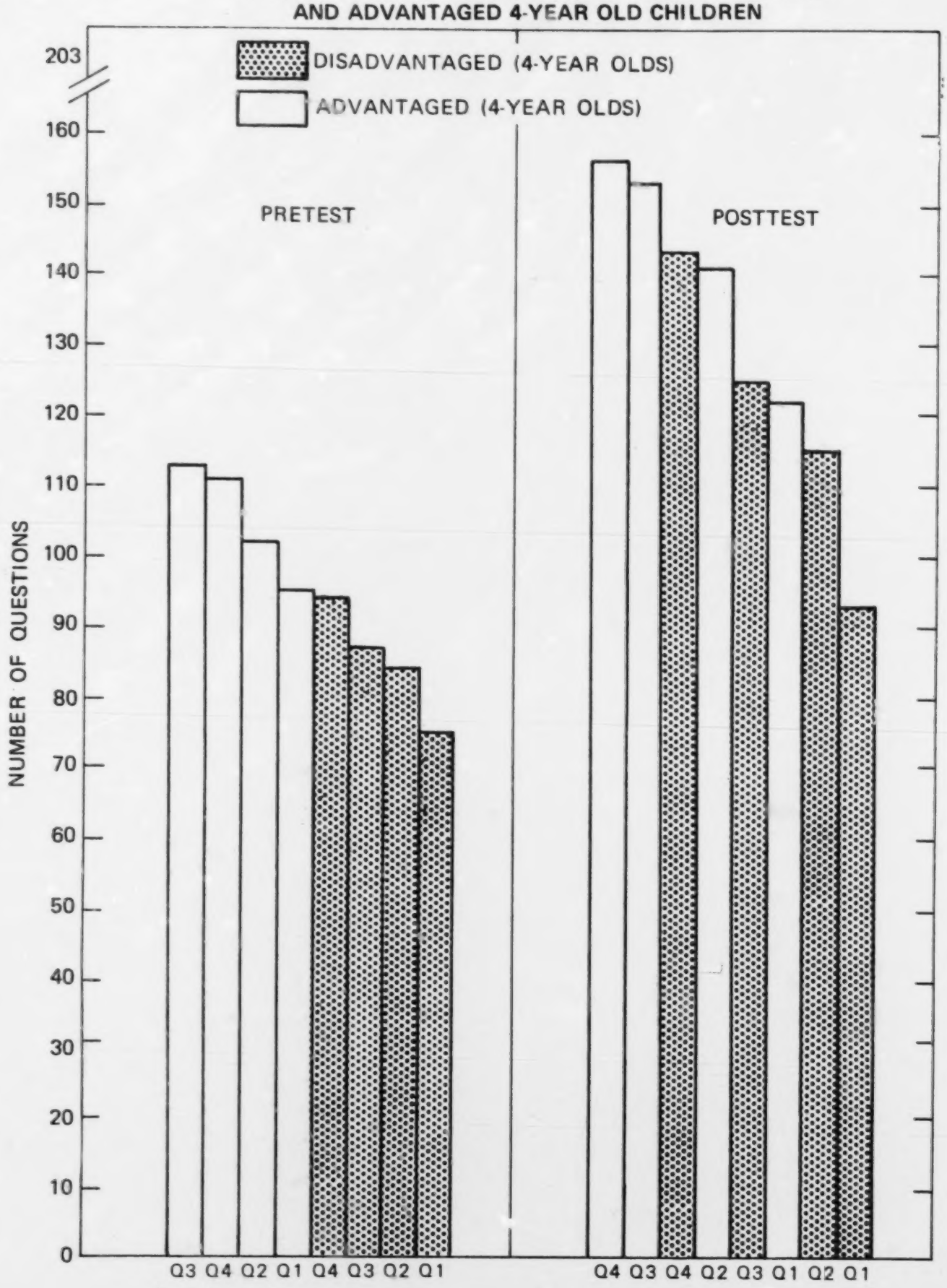
Insert Figure 3 about here

Figure 3 presents these data. It can be seen that at pretest time every group of advantaged children scored higher than every group of disadvantaged children. However, at posttest, the gains of Q3 and Q4 disadvantaged children resulted in a realignment; no longer were scores directly related to social class, but rather social class effects were clearly modified by amount of viewing. Disadvantaged children who often watched Sesame Street performed better on the measures of the show's goals than advantaged children who watched Sesame Street rarely or never.

The third result, also somewhat surprising, concerned the differences and lack of differences between at-home and at-Head Start disadvantaged children. Predictably, the scores of these groups differed at pretest, but there was no interaction between amount of viewing and home-Head Start status. Children at home gained about as much as children at school at each of the levels of amount of viewing.

Perhaps children at school were more readily distracted during viewing due to the group-viewing conditions and the availability of alternative sources of satisfaction in the classroom. As well, there was evidence that teachers used the show hour as an enrichment element in their program rather than as a central element in the curriculum. Follow-up activities were by no means universal. The other possibility, and the one we lean most to, is that Sesame

FIGURE 3
PRETEST AND POSTTEST SCORES OF DISADVANTAGED AND ADVANTAGED 4-YEAR OLD CHILDREN



Note how at pretest all advantaged groups do better than all disadvantaged groups. However, at post test, note how the high viewing disadvantaged groups surpass the lower viewing advantaged groups.

Street met one of the criteria it set for itself -- that it effectively taught at-home, preschool-aged children without dependence upon formal adult supportive roles. If television can be effective in this way, it suggests a number of uses in formal education via cassettes to help individualize instruction. It also suggests that a much greater drive is needed to ensure that children's TV programming is radically changed from its present lamentable status.

One interpretation which is not warranted and which has been put forward is that Sesame Street is an alternative to Head Start. Clearly the TV show is not an alternative in that its scope, goals, and functions are much more limited in comparison to those of Head Start. We have suddenly found ourselves in the position of supposedly advocating Sesame Street over Head Start. Our research in no way substantiates that position and nowhere have we ever advocated it. Political motivations seem to plague the interpretations of evaluations.

Post-Report Reactions

After the Sesame Street first year summative research report was made public, a critical technical reaction occurred. The argument concerned the assessment of amount of viewing as used in the study. This was a composite of four different measures. One of the measures was said to be suspect. The four measures were:

1. The posttest parent questionnaire in which a number of questions were asked about the viewing habits of the child.
2. The viewing record in which the parents of all encouraged at-home children and the teachers of all at-school encouraged children kept a daily record of amount of Sesame Street viewing.
3. The TV log in which, once a month, the parents of all at-home children circled the shows that their child watched that day.
4. The Sesame Street test in which all children at posttest were shown pictures of central characters on the show and asked if they could name or recognize them.

It was the last of these measures which created the problem. It was argued that this test was both a measure of viewing and a measure of learning. Of course we had no perfect measure of amount of viewing partly because, as Neilsen knows, one does not exist. It would be unnecessary to point out the deficiencies in the other three. We doubt that the Sesame Street test was the worst measure of amount of viewing, but on the surface it did present some problems of "confounding." So we ran the major analyses again using the first three measures separately and in combination as our indices of amount of viewing. The results were almost identical to those presented in the report except in one respect. If the Sesame Street test is eliminated, the pretest data show little

increase with amount of viewing. In general, those who viewed the show most did little better at pretest than those who viewed the show least. Thus, the very large increases seen at posttest disproportionately favoring the higher viewing groups became more readily interpretable.

We had been worried about the potential problem of assessing amount of viewing, so we had deliberately used four measures, and it is well that we did. The moral is that when a variable seems difficult to assess, try to use a number of different measures. Unfortunately the area of child psychology is laden with examples of research where this was not done--for example, research in achievement motivation, self-esteem, and anxiety. Many of the measures, say for anxiety, have but low relationship with other measures also ostensibly measuring anxiety. It comes back, of course, to our primitive state of knowledge about the conceptualization and measurement of some very important variables, and at least until this improves it would be well not to rely upon just one measure of a particular construct.

Some Concluding Remarks

We have not tried to present a comprehensive description of our summative research on Sesame Street. Rather we have tried to indicate those aspects of the research that were bases for making generalizations about research into the development of preschool-aged children. Incidentally, we have tried to suggest the need for child psychologists to become more in-

volved in, and attuned to, the needs of the world of educational evaluation. In fairness, we should also emphasize the need for educational evaluators to become more knowledgeable about the world of developmental research.

It is true that in large-scale field research you find you have to live with a diminution in precision. You hope that the less precise hammering of a large project at the gateway to knowledge will, in the long run, be as effective as the more precise tapping of smaller scale laboratory research. But the probability is that the two, in concert, might be the most effective.

Bibliography

Ball, Samuel and Bogatz, Gerry Ann. The First Year of Sesame Street: An Evaluation. Princeton, N.J.: Educational Testing Service, 1970.

Caro, Francis G. "Issues in the Evaluation of Social Programs," Review of Educational Research, 1971, Vol. 41, 2.

Freeman, H. E. "Strategy of Social Policy Research," In H. E. Freeman (Ed.), Social Welfare Forum. New York: Columbia University Press, 1963.

Rust, Lang. Attributes that Differentiate Boys' and Girls' Preferences for Materials in the Preschool Classroom: A Systems Design Approach. An unpublished doctoral dissertation, Teachers College, Columbia University, 1971.

Scriven, M. "The Methodology of Evaluation," AERA Monograph Series on Curriculum Evaluation, No. 1, Perspectives of Curriculum Evaluation, 1967.

Walsh, J. Anti-poverty R & D: Chicago Debacle Suggests Pitfalls Facing OEO. Science, 1969, 165, 1243-1245.